ESTRUCTURA SECRETA DE LA WEB. COMO ACCEDER A LAS FUENTES NO INDEXADAS.

Francisco Javier García Algarra. UNED/Telefónica Investigación y Desarrollo.[1]

Resumen

Internet plantea retos a la forma de trabajar de los historiadores. La inmensa cantidad de información que se produce y almacena está disponible para los investigadores donde y cuando quiera que la necesiten. Esta es una característica destacada de la era digital, pero no toda esta información se puede encontrar fácilmente. Los buscadores actuales solo indexan alrededor del 1% de los contenidos, por distintas razones técnicas. El 99% que queda fuera se denomina la web profunda y existen diferentes estrategias para vencer este obstáculo. Esta comunicación explica como encontrar documentos en la web profunda.

Abstract

Internet challenges the way of working of historians. The huge amount of information that produces and stores is available for researchers wherever and whenever they need it. This is a remarkable feature of the digital era, but not all this information is easy to find. Current search engines only index around 1% of contents, by different technical issues. The 99% that lies outside is called the deep web and there are different strategies to overcome this obstacle. This communication explains how to find documents inside the deep web.

Introducción

La documentación es la materia prima de la ciencia histórica y la eclosión de Internet desde finales del siglo XX ha introducido cambios radicales en su manejo. El volumen de información crece a una velocidad de vértigo y aparecen nuevas clases de documentos de naturaleza heterogénea y volátil, lo que complica su conservación y estudio. Además, el universo digital no es solo un nuevo medio de difusión de los datos, también es un escenario en el que sucede la historia.

1 Grupo de Investigación sobre Arte y Patrimonio Cultural de la Edad Contemporánea. Ingeniero del Grupo Telefónica desde 1991.

Los datos están ahí disponibles para quien quiera y sepa explotarlos, pero no es una labor sencilla encontrar lo que se necesita o valorar la calidad de lo que se encuentra. El orden y la confiabilidad de los archivos clásicos dejan paso a un torrente de documentación fragmentaria de origen incierto. El historiador debe saber manejarse en este medio y adquirir algunas de las competencias de un analista de información. Una de las mayores dificultades a las que debe enfrentarse es que la mayoría de los contenidos no aparece en los buscadores tradicionales como *Google* y *Bing*. Es lo que se denomina web profunda[2] y es importante ser conscientes de su existencia para poder aprovecharla.

Documentación en la era digital.

Según el último informe de *IDC*[3] el volumen de información generada cada año en el mundo crece a una tasa del 40% y en 2013 se produjeron 4.4 zettabytes[4]. Para tener una idea de lo que significa esta unidad de medida, un zettabyte podría almacenar un vídeo de alta calidad que durase 35 millones de años. Esos datos son los documentos de nuestra era: emisiones de televisión, entradas de blogs, *tweets*, publicaciones académicas, publicidad en todas sus formas, documentos oficiales, fotografías, transacciones bancarias, etc. Nunca la humanidad ha tenido tal capacidad de producir y almacenar datos y esto debería suponer una ventaja para los historiadores del futuro. Nuestras vidas se registran en tiempo real. Puede ser de forma legal, cuando hacemos la compra con nuestra tarjeta de crédito, consultamos una página web o compartimos las fotos de nuestras vacaciones pero también sabemos ahora que algunos gobiernos espían nuestro tráfico telefónico y nuestros correos, como así reveló el caso Snowden[5].

Sea cual sea el origen de la información, si se almacena podrá utilizarse para entender la historia de esta época. Su sobreabundancia no está exenta de inconvenientes, el mayor es la volatilidad de los nuevos tipos de documentación. Se suele distinguir entre

[2] Aunque existen antecedentes de uso del término, se considera que el artículo seminal sobre la web profunda (‘*deep web*’) apareció en 2001. Michael K. BERGMAN: *“White Paper: The Deep Web: Surfacing Hidden Value”, Journal of Electronic Publishing*, vol. 7, núm. 1 (2001).

[3] http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm. *IDC* es una compañía privada que desde 2005 publica informes anuales sobre el estado de Internet. Su estimación del volumen de datos se acepta como la más fiable en el mundo de la tecnología.

[4] Un zettabyte son10^{21} bytes. Un disco duro de ordenador, de 1 terabyte, contiene 10^{12} bytes, es decir la mil millonésima parte.

[5] Susan LANDAU: “*Making sense from Snowden”, IEEE Security & Privacy Magazine*, núm. 4 (2013), p. 5463.

información digital de nacimiento y aquella que ha sido digitalizada. Esta última la componen todas las publicaciones, manuscritos y material audiovisual que se han convertido a formato electrónico en las últimas dos décadas, y suele estar bien catalogada. La digital de nacimiento es mucho más frágil, si la consideramos en conjunto[6]. El cambio es constante y los contenidos no siempre se archivan. Lo que podía consultarse ayer hoy ya no existe, por motivos diversos que conducen a la desaparición de una porción importante de ese material.

El problema radica en los distintos grados de cristalización de los datos. En un extremo están las publicaciones académicas o los textos legales (p. ej. el BOE) que han dejado de publicarse en papel pero cuya razón de ser es la publicidad y perdurabilidad de sus contenidos. Muy próximos en cuanto a credibilidad están los catálogos *ad hoc* administrados por entidades como bibliotecas o universidades que ofrecen al investigador un alto grado de certeza sobre la fidelidad de la copia a la primera aparición. Iniciativas como el *Digital Object Identifier*[7] aseguran la calidad de estas fuentes.

En una segunda categoría se encuentran los archivos digitales de los grandes medios de comunicación. Los contenidos pueden presumirse bien conservados pero corren riesgo de desaparecer a medio y largo plazo por los avatares del mundo empresarial y no son inmunes a la censura[8]. Ofrecen un alto grado de verosimilitud respecto a la identidad del autor. Por ejemplo, un historiador de la economía que estudie la crisis del euro de 2012 y encuentre la entrada del blog del Premio Nobel Paul Krugman en la que este predecía la desaparición de la moneda en "cuestión de meses, no de años"[9] podrá estar bastante seguro de que era suya. Así lo acredita *The New York Times*, que es quien le pagaba por escribirlo y almacena todas sus contribuciones.

Las tiendas *on line* de productos culturales (libros, música, películas), disponen de excelente información de catálogo e infraestructura informática para mantenerla pero

---

[6] Meghan DOUGHERTY et al.: "*Researcher engagement with web archives: State of the art". Final Report for the JISC-funded project 'Researcher Engagement with Web Archives*, 2010.

[7] Administrado por la Digital DOI Foundation, una organización sin ánimo de lucro http://www.doi.org

[8] Un ejemplo es la eliminación de comentarios de los usuarios que contienen lenguaje malsonante o incitación a la violencia. Desde el punto de vista de un historiador de la cultura ese material puede ser mucho más valioso que la noticia en sí, sin embargo la fuente primaria se ha alterado. Un caso muy reciente, Kristen Hare: *"Guardian has deleted almost 500 comments from pro-Russia trolls",* 2014, http://bit.ly/1rzcxqO,

[9] Paul KRUGMAN: "*Eurodämmerung",* 2012, .http://krugman.blogs.nytimes.com/2012/05/13/eurodammerung-2

pueden eliminar ítems que ya no resultan de interés comercial o sufrir procesos de compra o quiebra.

Luego vienen los archivos de documentos compartidos como fotografías (*Instagram*, *Flickr*) o videos (*YouTube*, *Vimeo*) en los que la catalogación depende de las etiquetas que añaden los usuarios, no siempre acertada. *Wikipedia* puede incluirse en este apartado puesto que la teórica obligación de citar fuentes externas no se cumple en todos los casos, aunque permite de forma sencilla consultar las modificaciones de un artículo.

A partir de aquí la volatilidad crece de manera notable. Los mensajes en redes sociales (*Twitter*, *Facebook*, *LinkedIn*) se pueden consultar de manera ordenada porque los propietarios de estos sitios así lo estiman conveniente, pero no hay ninguna certeza sobre la identidad de los autores. Un estudio de 2102 encontró que el 11% de esos contenidos se pierden en el primer año y a partir del segundo desparecen a una tasa del 7%.[10]

La conservación de mensajes en foros web depende mucho del azar, pero no suelen sobrevivir al sitio que originalmente los cobijó y equivalen a notas manuscritas clavadas con chinchetas en un tablero de corcho.

Si pasamos a la información protegida por el secreto de las comunicaciones la situación es mucho más incierta. El correo electrónico ha demostrado ser perdurable y se acepta como prueba judicial. Imaginemos lo que pueden valer los contenidos de servicios como *Gmail* o *Hotmail* que usan millones de usuarios. Aunque en teoría son privadas, no sabemos lo que sucede con las conversaciones por sistemas de mensajería instantánea (*WhatsApp*), ni con las conferencias o videoconferencias (*Skype*, *Google Hangout*). La sospecha generalizada es que hay agencias de inteligencia que tienen acceso ellas. Otros muchos datos se pierden directamente porque no se almacenan.

En resumen, la información digital difiere mucho en cuanto a su grado de cristalización, catalogación y perdurabilidad, pero toda investigación suele empezar por el mismo punto, el buscador (*Google* y en mucha menor medida *Bing* y *Yahoo*[11]), convertido en la biblioteca de Babel de nuestro tiempo. Lo que muchos investigadores no conocen es

[10] Hany M. SALAHELDEEN, Michael L. NELSON: "*Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?",* arXiv:1209.3026, 2012.

[11] Según *StatCounter*, en mayo de 2014 *Google* acaparó el 88,5% de las búsquedas, por el 4,4% de *Bing* y el 3,5% de *Yahoo*. http://gs.statcounter.com/#search_engine-ww-monthly-201305-201405

que la gran mayoría de los datos de Internet no pueden encontrarse con los buscadores porque no están indexados, permanecen en la web profunda.

La web profunda, concepto y dimensión.

Los buscadores trabajan por una aproximación que puede considerarse de fuerza bruta, ideada a finales del siglo XX, cuando la mayor parte de los contenidos eran estáticos. Internet es muy anterior a la aparición de la *World Wide Web*[12], pero fue la popularización de este concepto, desarrollado en el *CERN*[13] por Tim Berners-Lee[14], el que la convirtió en un fenómeno de masas en los años 90. La idea básica de Berners-Lee era sencilla, almacenar los documentos en un formato de marcas especial (HTML[15]) que permitiera incluir enlaces hacia cualquier otro documento de la red. Para conseguirlo desarrolló el primer servidor web y el primer navegador.

Un fichero HTML contiene etiquetas de formato que indican al navegador como presentar el contenido, pero en esencia puede considerarse un fichero de texto estático. Por ejemplo, el comienzo del texto de Berners-Lee convertido a HTML que acabamos de citar tiene este aspecto:

```
<h1>Information Management: A Proposal</h1>
<address>
  Tim Berners-Lee, CERN <br>
  March 1989, May 1990
</address>

<p>This proposal concerns the management of general information
about accelerators and experiments at CERN. It discusses the
problems of loss of information about complex evolving systems
and derives a solution based on a distributed hypertext
system.</p>
```

[12] Véase Andreu VEÀ: *Como creamos internet*, Península, 2013. Su sitio web personal contiene material inédito muy interesante sobre esta creación colectiva y los que la hicieron posible http://comocreamosinternet.com/

[13] *Conseil Européen pour la Recherche Nucléaire*, posiblemente el organismo científico paneuropeo de mayor prestigio. cern.ch

[14] Tim BERNERS-LEE: "*Information Management: A Proposal"*, 1989. Puede consultarse en la siguiente dirección: http://www.w3.org/History/1989/proposal.html. Este documento es un excelente ejemplo del problema de la perdurabilidad de las fuentes digitales. El original era un fichero almacenado en un formato ya obsoleto que se distribuyó electrónicamente pero nunca se publicó impreso. El consorcio WWW conserva la primera versión del original convertida a HTML. Es un acto de fe del investigador creer la afirmación "*The text has not been changed, even to correct errors such as misnumbered figures or unfinished references*". No se conserva el original del acta fundacional de la web.

[15] *Hypert Text Markup Language*. El concepto de hipertexto es anterior y existían ya otros lenguajes similares, pero a nadie se le había ocurrido utilizarlos para enlazar documentos de cualquier ordenador conectado a la red.

```
<p align="center"><img src="Image1.gif" width="512" height="458"
alt="A circles and arrows diagram relating concepts discussed in
the paper"
border="0"></p>
```

Los buscadores recorren los servidores de la WWW y descargan todas las páginas públicas en sus inmensos almacenes de datos. Cuando el usuario lanza una petición, se desencadena una serie de procedimientos internos, muy complejos de describir para el alcance de esta comunicación, que localizan todas las apariciones de la cadena de texto solicitada en el almacén del buscador. Es importante entender esta manera de trabajar, la búsqueda no se hace rastreando toda la red en el momento en que se lanza, sino sobre la copia que el buscador ha ido construyendo poco a poco. Si pedimos a *Google* o a *Bing* que encuentren la cadena de texto "*accelerators and experiments at CERN*" devolverán las direcciones de todos los documentos de su almacén en los que aparece, junto con las direcciones (URL[16]) desde las que se descargaron. La magia se produce porque al hacer *click* sobre el hiperenlace el navegador lleva a la página original.

Esta solución resultaba muy conveniente con las páginas web de los años 90. Casi todas eran ficheros HTML y experimentaban pocos cambios, de manera que la probabilidad de que la copia en el almacén del buscador y el original coincidieran era alta. Sin embargo, la tecnología evolucionó y comenzaron a aparecer páginas dinámicas. Estas se generan al momento para cubrir una necesidad inmediata y después se desvanecen o cambian. Con dos ejemplos se entenderá mejor. La página principal de un periódico digital está variando de forma casi continua, a medida que se producen nuevas noticias. Esto no se hace modificando un fichero a mano, sino creando el contenido en ese instante mediante programas sofisticados, aunque lo que finalmente llega a nuestro navegador es HTML. Es imposible para el buscador mantener una copia fiel de todos los cambios, hay información que no puede indexar aunque de hecho sea accesible para cualquiera que consulta el periódico. Otro ejemplo cotidiano es una página web que permita consultar horarios y precios de vuelos. Cada vez que hacemos una petición se genera HTML a partir de una colección interna de datos y todos tenemos la experiencia de la rapidez con la que cambian los resultados. El buscador tampoco puede seguir el ritmo de estas páginas.

[16] *Uniform resource locator*, compuesto por el nombre del servidor, el directorio y el nombre del fichero. Por ejemplo http://home.web.cern.ch/about/computing.

¿Qué ocurre cuando el acceso está restringido por usuario y clave? Los buscadores se encuentran con un obstáculo insalvable, no tienen acceso más allá de la página inicial y aclaramos que por fortuna, ya que así no pueden entrar en nuestras cuentas bancarios o historiales médicos.

No es posible indexar los contenidos de las bases de datos aunque su acceso sea público, el buscador solo accede a la interfaz web desde la que el usuario lanza las consultas. La Hemeroteca Digital de la Biblioteca Nacional de España[17] es un buen ejemplo, dispone de un buscador propio muy potente para consultar su gran colección de prensa histórica digitalizada, pero Google no tiene forma de indexar esos contenidos. Su homóloga francesa gestiona *Gallica*[18] que contiene cinco millones de ítems digitalizados y presenta la misma limitación.

Otra complicación que ha aparecido en los últimos años es el crecimiento explosivo de los contenidos audiovisuales. Un video necesita un espacio de almacenamiento muy superior al de cualquier texto, es virtualmente imposible que los buscadores actuales puedan copiar localmente todo ese volumen, y eso sin entrar en los retos técnicos que plantea la localización de la información en este tipo de soportes.

Añadamos a esta lista la de los sitios legales que no quieren aparecer en los buscadores y que pueden hacerlo incluyendo un pequeño fichero y la de los del lado más oscuro de Internet que hacen todo lo posible para permanecer ocultos[19].

El conjunto de todos los contenidos no indexados es lo que se conoce como la web profunda, y la única característica que comparten en que no pueden encontrarse con un buscador clásico. La metáfora es de Michael K. Bergman en el artículo de 2001 antes citado, los buscadores son como barcos pesqueros que echan sus redes de arrastre en la capa más superficial del mar de los datos, pero las profundidades abisales están fuera de su alcance.

El tamaño de la web profunda es muy difícil de medir porque no existe un almacén único de datos. La cifra más repetida es que supone un 99,7% de toda la información, esa estimación es la que hizo Bergman hace 13 años durante los cuales la red ha evolucionado muchísimo. Los trabajos más recientes coinciden en que supone más del

[17] http://www.bne.es/es/Catalogos/HemerotecaDigital/
[18] http://gallica.bnf.fr
[19] Las referencias a la web profunda en prensa suelen centrarse en este segmento, atractivo como todo lo prohibido. No obstante, conviene aclarar que el término no tiene ningún matiz peyorativo, ni se limita a sórdidos sitios clandestinos que trafican con armas, drogas o seres humanos.

99% del total, y eso permite hacerse una idea de la magnitud del problema[20].

El acceso a los contenidos de la web profunda es un campo floreciente en la comunidad académica relacionada con la tecnología, pero apenas mencionado por los investigadores en humanidades[21]. Es importante saber que el obstáculo existe, conocer las estrategias para superarlo y adquirir las habilidades necesarias para adaptar nuestro trabajo a un entorno en permanente transformación. No se trata de que el historiador se transforme en un científico de datos (*data scientist*) pero sí de que entienda los métodos y el lenguaje de los profesionales de la tecnología para poder establecer una colaboración fructífera y equilibrada[22].

Como aprovechar la riqueza de la web profunda

Es un hecho que el 99% de la información de Internet no se puede encontrar desde nuestro buscador favorito. Eso significa que existe un inmenso territorio por explorar. Si nos limitamos al 1% de acceso más cómodo estaremos compitiendo con otros muchos peces en la superficie del mar y será difícil realizar hallazgos originales, pero la situación no es muy diferente de la que se producía en la investigación tradicional.

[20] Bin HE et al : "*Accessing the deep web"*, *Communications of the ACM*, vol. 50, núm. 5 (2005), pp. 94-101. Denis SHESTAKOV, Tapio SALAKOSKI: "*On estimating the scale of national deep Web"*, *Database and Expert Systems Applications*. Springer Berlin Heidelberg, 2007 pp. 780-789. Jayan MADHAVAN et al.: "*Harnessing the Deep Web: Present and Future",* arXiv: 0909.1785, 2005. Ritu KHARE, Yaun AN, Il-Yeol SONG: *"Understanding deep web search interfaces: a survey", ACM SIGMOD Record*, vol. 39, núm. 1 (2010), pp. 33-40. Debido a que los buscadores utilizan distintas estrategias de muestreo y almacenamiento, hay contenidos que pueden aparecer solo en una parte de ellos pero no en los otros. Esto llevaría a establecer matices en la dificultad de la búsqueda entre aquellos sitios que aparecen en todos y aquellos que no están en ninguno, pero debido a la preponderancia de *Google* en el mercado, la definición actual de web profunda es prácticamente equivalente a la del conjunto de recursos no indexados por este buscador.

[21] Véase como excepción Elisabeth YAKEL: "*Searching and Seeking in the Deep Web: Primary Sources on the Internet"*, en *Working in the Archives: Practical Research Methods for Rhetoric and Composition*, 2010, p. 102-118.

[22] Como en todos los ámbitos de la actividad humana, si existe un nicho de mercado sin explotar alguien lo ocupará. En fecha muy reciente la prensa mundial se hizo eco de que, según un algoritmo, la persona más influyente de la historia fue el botánico sueco Carl Linnaeus, resultado chocante para cualquier historiador. Si se lee el borrador publicado por los autores, lo que la investigación determina son las páginas de personajes que tienen un mayor índice de centralidad en *Wikipedia*, en el sentido de la teoría de redes complejas. Aplicaron el mismo procedimiento que usa *Google* para ordenar sus resultados. En ningún caso afirman que eso se traduzca en un imaginario coeficiente de importancia histórica. La noticia se propagó sin el menor sentido crítico y seguramente se repetirá por mucho tiempo al haber alcanzado el nivel de "meme" o fenómeno viral. Los investigadores en humanidades pueden caer en el error de reproducirla por un miedo atávico a llevar la contraria a un artículo lleno de fórmulas matemáticas. Eom YOUNG-HO et al.: "*Interactions of cultures and top people of Wikipedia from ranking of 24 language editions"*, arXiv:1405.7183, 2014.

Siempre ha habido archivos por explotar y colecciones documentales que no eran más que una acumulación de legajos o expedientes administrativos sin orden. La labor creativa en historia radica en la elaboración de un discurso a partir de esos datos en bruto y de sus relaciones. No hay una receta única para acceder a todos los contenidos de la web profunda[23] pero sí se pueden seguir unas pautas en función del tipo de barrera que lo impida.

Existen catálogos generales que recogen información extraída de fuentes no indexables. Algunos son muy antiguos como *dmoz*[24], un directorio de páginas catalogadas a mano por millones de usuarios, que recuerda como era el mundo digital antes de la aparición de *Google*. El más conocido es *WorldCat*, capaz de buscar en dos mil millones de registros de bibliotecas y archivos de todo el mundo y que es un buen inicio para empezar a explorar. Otros sitios que fusionan datos de diversa procedencia funcionan al modo de enciclopedias (*Infoplease.com* o *libraryspot.com*) pero resultan más interesantes los especializados en una disciplina o tema concreto. La Universidad de Idaho mantiene un catálogo de fuentes primarias organizado por geografía, con enlaces a cinco mil archivos[25]. En España disponemos del portal *PARES*[26] que recoge datos de múltiples instituciones y es una referencia obligada. Son muy numerosos los especializados, podemos citar como ejemplo reciente, la base documental de arquitectura industrial del siglo XX[27] mantenida por la *ETSAM* que se ha confeccionado con todas las publicaciones aparecidas entre 1940 y 1981.

Las revistas académicas conforman un conjunto de máximo interés para el historiador y en gran medida forman parte de la web profunda porque están protegidas por un muro de pago. Aunque las universidades permiten el acceso a algunas de ellas, es imposible estar suscrito a todas por el coste tan elevado que supone. Contra esta limitación funciona una aproximación en dos pasos, la primera es la localización de la referencia en un catálogo general como *Google Scholar*, *CiteSeer* o *Dialnet*. Si existen copias públicas del documento en PDF u otro formato estamos hablando de web superficial, pero es muy habitual localizar solo resúmenes o la primera página, y esta limitación se

[23] Si fuese posible automatizar el acceso a toda la web profunda los buscadores lo harían. De hecho invierten grandes sumas en ampliar su alcance. Véase Jayant MADHAVAN et al.: "*Google's deep web crawl", Proceedings of the VLDB Endowment*, vol. 1, núm. 2 (2008), pp. 1241-1252.

[24] http://www.dmoz.org

[25] http://webpages.uidaho.edu/special-collections/other.repositories.html

[26] http://pares.mcu.es/

[27] http://www.arquitecturaeindustria.org

debe casi siempre al convenio de publicación. En este caso, conviene formar parte de comunidades de investigadores como *Academia.edu* o *Researchgate.org*. Es habitual que el contrato de cesión de derechos consienta compartir copias a título personal con colegas para fines académicos. Estas comunidades permiten solicitar a otros usuarios la publicación restringida del documento sin violar esas condiciones. Su uso es muy habitual en ciencia y tecnología y menos en humanidades. Otros servicios de pago como *JSTOR*[28] dan acceso gratuito a los contenidos de publicaciones cuando ha pasado un tiempo desde su aparición.

Una práctica en la que también difieren las costumbres de la ciencia y las humanidades es la publicación de borradores (*eprints*) que responde a una necesidad del primer grupo. Las novedades tienen un periodo de vigencia muy breve y es vital publicar cuanto antes. Esto no casa bien con los plazos de revisión por pares que son similares en cualquier publicación académica. Nació así *arXiv*[29] en 1991, una web para físicos que permitía publicar resultados antes de enviarlos a una revista. Con el tiempo su uso se ha extendido a otras disciplinas científicas aunque no a las ciencias sociales, con excepción de la economía. Siguiendo este ejemplo nació en 1994 *SSRN* (*Social Science Research Network*), que publica borradores en múltiples categorías. Este tipo de sitios permiten descargar legalmente los contenidos tal y como eran antes de someterse al proceso de revisión.

El análisis del contenido de las redes sociales es un terreno muy activo[30]. Se presta a procedimientos automáticos y trabaja con una muestra muy numerosa de la población, aunque no aleatoria. La red más estudiada es *Twitter*, porque sus contenidos son públicos y dispone de un mecanismo para descargar la información de una forma sencilla. Los pioneros de este análisis procedían del campo de la ciencia de redes, pero cada día se han ido sumando expertos de más disciplinas. Es una fuente que se empieza a explotar desde las humanidades[31].

Al principio de esta comunicación hemos afirmado que la red no es solo crónica de la

---

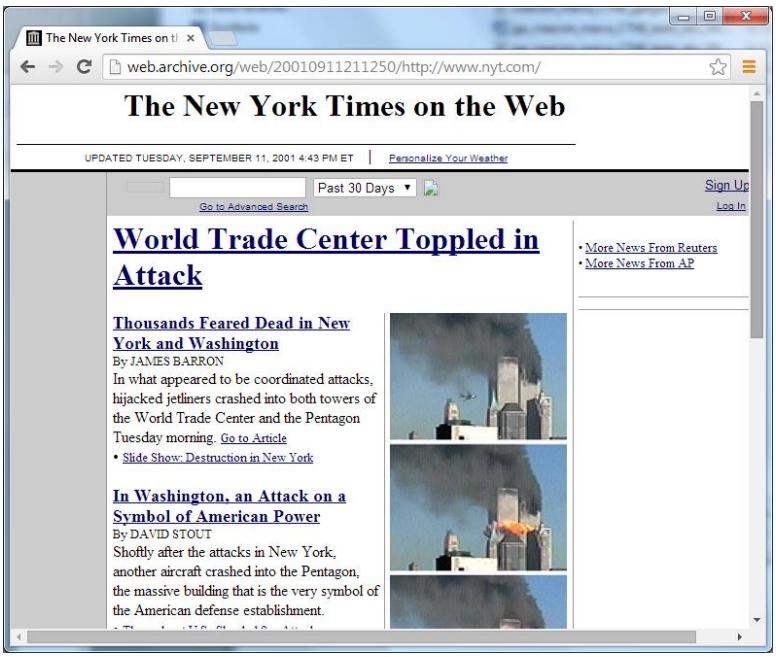
[28] http://www.jstor.org

[29] En la actualidad lo gestiona la Universidad de Cornell http://arxiv.org

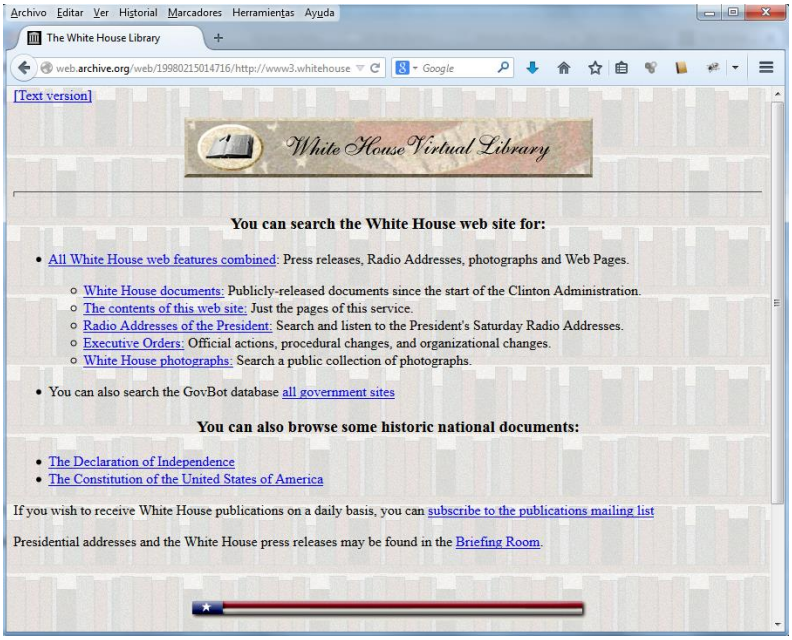
[30] Véase la revista Social Networks. An International Journal of Structural Analysis,

[31] Sean ADAY, et al.: "*New Media and Conflict after the Arab Spring"*. Washington DC: United States Institute of Peace, 2012. Narseo VALLINA-RODRIGUEZ et al: "*Los twindignados: The rise of the indignados movement on twitter", Privacy, Security, Risk and Trust* (PASSAT), International Conference on Social Computing (SocialCom), 2012, pp. 496-501. Ismael PEÑA-LÓPEZ, Mariluz CONGOSTO, Pablo ARAGÓN: "*Spanish Indignados and the evolution of 15M: towards networked para-institutions"*. En Congreso Internacional Internet, Derecho y Política, 2013.

actualidad, sino escenario donde discurre la historia contemporánea. El escenario evoluciona día a día pero los documentos nacen en momento dado y necesitamos conocerlos tal y como eran en ese instante. Hay portadas de periódico que se convierten en icono, como las que anunciaron el comienzo de la Segunda Guerra Mundial, y conservan la intensidad dramática del acontecimiento. Pero, ¿qué ocurre con los grandes sucesos de la era digital?



La imagen anterior es la página web del *New York Times*, el 11 de septiembre de 2001 a las 21:12, hora local de aquel día. Tiene un aspecto primitivo, incluso *amateur*, si la comparamos con cualquier medio digital de 2014, pero posee el enorme valor de conservar un instante histórico. La página oficial de la Casa Blanca aparecía así el 15 de febrero de 1998. Parece el ejercicio de un colegial pero así era la estética imperante.



Estos dos ejemplos se han obtenido del mejor recurso para la navegación en el tiempo, el *Internet Archive*[32]. Este gigantesco depósito de recursos empezó a funcionar en 1996 y guarda material muy diverso: colecciones de libros publicados hace 50 o más años cedidos por bibliotecas, películas, fotografías, revistas y archivos de audio. Una de sus secciones se dedica a conservar materiales de la propia historia de la red[33], pero la más valiosa es la *Wayback Machine*. El archivo realiza descargas diarias de cientos de miles de páginas y las almacena ordenadamente. En la actualidad la cifra asciende a 415 mil millones de capturas. Este material pertenece en su mayor parte a la web profunda, ya no existe para los buscadores comerciales pero sí para la inmensa base de datos de este servicio gratuito. Podemos navegar a cualquier fecha del pasado y recuperar la información tal y como era, lo que ofrece unas posibilidades inmensas al historiador. Además dispone de un método de descarga automático que facilita la automatización del análisis de los contenidos.

Existen otras iniciativas similares al *Internet Archive*, aunque de ámbito más limitado. *PANDORA*[34] es el archivo web de Australia, y uno de los pioneros puesto que empezó a funcionar en 1996. *UK Web Archive*[35] desarrolla una labor similar a la de *Wayback*

[32] La dirección es archive.org, no debe confundirse con el antes aludido arXiv.org. Los contenidos están replicados por seguridad en la *Bibliotheca Alexandrina* www.bibalex.eg.
[33] archive-it.org
[34] pandora.nla.gov.au
[35] www.webarchive.org.uk

Machine, centrándose en las páginas británicas. La Biblioteca Nacional de España y Red.es están construyendo el Archivo de la web española[36], y *PADICAT*[37] mantiene el histórico de unas 65.000 páginas catalanas. La mayoría de estos organismos pertenecen al *International Internet Preservation Consortium*[38].

El servicio *archive.today* permite al usuario introducir una URL y almacenar de forma permanente su contenido, aunque se trate de una página dinámica. Hay también proyectos sectoriales como el *CyberCemetery*[39], que guarda todos los materiales digitales de los organismos federales de Estados Unidos que cesan en su actividad y se clausuran.

La situación se vuelve más complicada cuando se buscan fuentes audiovisuales. El mayor volumen de datos que se genera a diario tiene estos formatos y resulta complicado almacenarlo en su totalidad. Aunque los procedimientos de extracción automática de datos desde archivos de video o las búsquedas por imagen han avanzado mucho, se sigue dependiendo en gran medida de las anotaciones que describen el contenido. No cabe duda de que estos objetos culturales son imprescindibles para entender nuestro tiempo. Cualquier evento puede quedar registrado por las cámaras de miles de millones de teléfonos móviles delos ciudadanos convertidos en reporteros[40]. A pesar de la desventaja de su gran tamaño los archivos de video y audio son información mucho más cristalizada que la de una página web. Aunque se pueden editar y modificar, existen herramientas para detectar estas manipulaciones, por lo que son fuentes primarias de gran calidad. El ejemplar más célebre y estudiado de esta categoría es la película super 8 en la que Abraham Zapruder captó el magnicidio de Kennedy. De no haber existido, la reconstrucción de los hechos y el relato histórico habrían resultado más incompletos. La historia contemporánea no podrá hacerse dejando de lado la expresión más genuina de nuestra civilización, habrá que emplear mucho tiempo en ver películas como antes se necesitaban días para revisar microfilms.

Para finalizar, hay que hacer mención a los rincones tenebrosos de la web profunda[41],

---

[36] www.bne.es/es/LaBNE/ArchivoWeb/index.html
[37] www.padi.cat
[38] netpreserve.org
[39] govinfo.library.unt.edu/default.htm
[40] Stuart ALLAN, Einar THORSEN: "*Citizen journalism: Global perspectives",* Peter Lang, 2009. Mark DEUZE: "*The changing context of news work: Liquid journalism for a monitorial citizenry"*, *International Journal of Communication*, vol. 2 (2009), p. 18.
[41] Symon AKED et al.: "*Determining What Characteristics Constitute a Darknet*, en *Proceedings of the*

que son los que más avivan la curiosidad de los que se acercan a este tema por primera vez. Internet es un reflejo de la sociedad y como en ella hay zonas muy oscuras. Existe un dominio de red *.onion* que permite no revelar el nombre del sitio web. Por el contrario, se publica con una críptica combinación alfanumérica, por ejemplo *8724G8723GJYGE30.onion*. Hay también un programa para navegar de forma anónima por esta red, denominado TOR. Estos sitios no se indexan en los buscadores tradicionales, pero sí mantienen directorios parciales que intercambian entre ellos. Se pueden encontrar también catálogos de sitios *.onion* en la web superficial, el más conocido es *The Hidden Wiki*[42]. En este anonimato se desarrollan actividades de todo tipo ya sean lícitas o ilícitas. No es una zona habitual para la investigación histórica, pero puede ser una fuente interesante para determinados fenómenos. Los documentos filtrados por *WikiLeaks* se movían en este estrato de la red mucho antes de hacerse públicos[43] y se publican miles de supuestos archivos secretos de gobiernos y empresas. El problema es verificar la autenticidad de la información.

Conclusiones

Gran parte de los documentos que permitirán estudiar la historia contemporánea nacen y se almacenan en Internet pero solo un porcentaje muy reducido está indexado en los buscadores. Es imprescindible que los historiadores conozcan este hecho y sepan cómo localizar las fuentes que se localizan en la web profunda. Los contenidos de la red son volátiles, por ello resultan necesarios los archivos de páginas web que deben explotarse como almacenes de documentación valiosa. Su mantenimiento financiero depende de su utilidad como herramienta de investigación, de manera que es nuestra responsabilidad aprovechar su existencia y cuidar de ellos como de los archivos documentales clásicos.

11th Australian Information Security Management Conference", 2013.

[42] www.thehiddenwiki.net

[43] Neil COOKE, Lee GILLAM: "*Clowns, Crowds, and Clouds: A Cross-Enterprise Approach to Detecting Information Leakage Without Leaking Information"*, en *Cloud Computing for Enterprise Architectures.* Springer, 2011, pp. 301-322.